Detecting Virus Exposure During the Pre-Symptomatic Incubation Period Using Physiological Data

Lauren Milechin[1], Shakti Davis[1], Tejash Patel[1], Mark Hernandez[1], Greg Ciccarelli[1], Steven Schwartz[1], Lisa Hensley[2,#], Arthur Goff[2], John Trefry[2], Catherine Cabrera[1], Jack Fleischman[1], Albert Reuther[1], Franco Rossi[2], Anna Honko[2,#], William Pratt[2], Albert Swiston[1,*]

[1] Massachusetts Institute of Technology Lincoln Laboratory, Lexington MA

[2] US Army Medical Research Institute of Infectious Diseases, Ft. Detrick MD

Currently at the Integrated Research Facility, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Ft. Detrick MD
* To whom correspondence should be addressed: albert.swiston@ll.mit.edu

Keywords: Ebola, Marburg, filovirus, hemorrhagic fever, machine learning, random forests, physiology, incubation period, pre-symptomatic, pre-fever, early infection detection, early warning system, telemetry

Abstract

Early pathogen exposure detection allows better patient care and faster implementation of public health measures (patient isolation, contact tracing). Existing exposure detection most frequently relies on overt clinical symptoms, namely fever, during the infectious prodromal period. We have developed a robust machine learning method to better detect asymptomatic states during the incubation period using subtle, sub-clinical physiological markers. Using high-resolution physiological data from non-human primate studies of Ebola and Marburg viruses, we pre-processed the data to reduce short-term variability and normalize diurnal variations, then provided these to a supervised random forest classification algorithm. In most subjects detection is achieved well before the onset of fever; subject cross-validation lead to 52±14h mean early detection (at >0.90 area under the receiver-operating characteristic curve). Cross-cohort tests across pathogens and exposure routes also lead to successful early detection (28±16h and 43±22h, respectively). We discuss which physiological indicators are most informative for early detection and options for extending this capability to lower data resolution and wearable, non-invasive sensors.

Introduction

We have developed a method for assessing viral exposure based solely on host physiological signals, in contrast to conventional diagnostics based on fever or biomolecules [1] of the pathogen itself or the host's immune response. Early warning of pathogen exposure has many advantages: earlier patient care increases the probability of a positive prognosis [2-5] and faster public health measure deployment, such as patient isolation and contact tracing [6-8], which reduces transmission [9]. Following pathogen exposure, there exists a "pre-symptomatic" incubation phase where overt clinical symptoms are not yet present [10]. This incubation phase can vary from days to years depending on the virus [11,12], and is reported to be 3-25 days for many hemorrhagic fevers [3,4,13,14]. Following this incubation phase, the prodromal period is marked by non-specific symptoms such as fever, rash, loss of appetite, and hypersomnia [10]. Figure 1 shows a conceptual model of the probability of infection detection P_d during different post-exposure periods (incubation, prodrome, and virus-specific symptoms) for current specific and non-specific (i.e., symptoms-based) diagnostics. We also include what may be considered an "ideal" sensor system capable of detecting viral exposure even during the earliest incubation period. We hypothesize that quantifiable abnormalities (versus a diurnal baseline, for instance) in high-resolution physiological signals, such as those from electrocardiography, hemodynamics, and temperature, *before* overt clinical signs could be a basis for the ideal signal in Figure 1, thereby providing advanced warning (the early warning time, Δt) of on-coming illness.

In addition to characteristic clinical presentations, most infectious disease diagnosis is based upon identification of pathogen-specific molecular signatures (via culture, PCR/RT-PCR or sequencing for DNA or RNA, or immunocapture assays for antigen or antibody) in a relevant biological fluid [10,15-22]. Exciting new approaches allowed by high-throughput sequencing have shown the promise of pre-symptomatic detection using genomic [23,24] or transcriptional [25] expression profiles in the host [26]. However, these approaches suffer from often prohibitively steep logistic burdens and associated costs (cold chain storage, equipment requirements, extremely qualified

operators, serial sampling): indeed, most infections presented clinically are never definitively determined etiologically, much less serially sampled. Furthermore, molecular diagnostics are rarely used until patient self-reporting and presentation of overt clinical symptoms, such as fever. Past physiological signal-based early infection detection work has been heavily focused on bacterial infection [27-32], and largely centered upon higher time resolution analysis of body core temperature [32,33], advanced analyses of strongly-confounded signals such as heart rate variability [28-30] or social dynamics [34], or sensor data fusion from already symptomatic (febrile) viral-infected individuals [35]. While great progress has been made in developing techniques for signal-based early warning of bacterial infections, we are unaware of any effort in extending these techniques to possibly life-threatening viral infections.

Electronics miniaturization has led to a wave of wearable sensing technologies for health monitoring [36], and increasingly more processing power is available to consumers to make meaningful use of these collected data [37]. Inspired by these developments, we envision a low-profile, robust, wearable, personalized and multi-modal physiological monitoring system persistently measuring signals capable of sensitive pathogen infection detection. Such a system could cue the use of highly specific (but expensive) diagnostic tests, prompt low-regret responses such as patient isolation and observation, or advise clinicians of fulminant complications in already compromised patients.

We use high-resolution physiological data from non-human primates (NHPs) exposed via intramuscular (IM) or aerosol routes to either of two viral hemorrhagic fevers (Ebola virus [EBOV] and Marburg virus [MARV]) to build this novel high sensitivity, low etiological specificity (that is, not informative of particular pathogens) processing and detection algorithm. Data is normalized to remove diurnal rhythms, aggregated to reduce short-term fluctuations, and then provided to a supervised binary classification (exposed and unexposed) machine learning algorithm as illustrated in Figure 2(a). We tested and compared several methods; RFs had the best positive predictive value (discussed below) and were chosen for the rest of our analysis. RFs were also chosen for their robustness against feature-rich and noisy data while minimizing over-fitting. Random forests are grown (trained) at two post-exposure stages, allowing the

algorithms to adapt to physiological changes between incubation and prodromal phases. One RF is trained using pre-fever physiological data and the other using post-fever data. Both RFs include pre-exposure data to build the unexposed class. Subject data is separated into training and testing sets, and every testing subject's data is provided to the RF model for an exposure prediction every 30 min. After using binary integration and a constant false alarm thresholding approach to further reduce false alarms, mean exposure declaration times are found to range from 21h (for EBOV) to 69h (for MARV) before the onset of fever defined as 1.5^{o}C above a diurnal baseline [38] sustained for two hours. Figure 2(b) shows a block diagram of the declaration process. We note that all physiological data is given to our algorithm without regard to exposure or fever status; in other words, our approach does not require information on exposure or fever times for successful classification and detection.

Implementing this type of early-warning algorithm could save lives of health care workers, military service members, patients, and other susceptible individuals. During the 2014 West Africa Ebola outbreak, for instance, health care workers at higher risk of viral exposure could have been monitored persistently for the earliest possible indications of viral exposure. More commonly, patients in post-operative or critical care units could be monitored for infection and treated well before clinical symptoms, viremia/bacteremia, or septic shock [39]. In future, etiologically-specific iterations of this approach, knowledge of causative pathogens could inform very early therapeutic intervention. Furthermore, using very feature sparse datasets, such as those that could be collected using wearable sensor platforms, would enable this technique to be implemented in non-ideal clinical, athletic, and military environments. Transitioning this technology to these contexts is the focus of ongoing work.

Results

We use high resolution physiological data collected during previously conducted natural history studies (presently unpublished) at the United States Army Medical Research Institute of Infectious Diseases (USAMRIID) to build a binary classification random forest machine learning model [40] for detecting whether an animal had been

exposed to a viral hemorrhagic fever virus (either Ebola or Marburg virus). Supervised machine learning algorithms observe characteristics in data that belong to pre-determined classes, then place new, unseen data into the appropriate class based on similar characteristics. Here, we define pre- and post-exposure as the two classes since "infection" is not a discrete event.

We experimented with several classification methods, including Naïve Bayes[41], k-Nearest Neighbors [42], and random forests (RFs), and compared each across sensitivity, specificity, and early warning time metrics. All classifiers had positive predictive values, yet chose RFs for several reasons (results for other classifiers are found in Supplementary Figure 1). Most importantly, Random Forests require no assumptions about the statistical independence of features, which is critically useful given highly correlated physiological feature sets. RFs also allow for the calculation of quantitative feature performance; this also facilitates post-hoc comparison to the known viral pathology sequence to mechanistically understand why these physiological anomalies are present. Furthermore, the most discriminating features can be selectively chosen to re-grow forests and allow for better algorithm performance with fewer feature inputs. Next, a collection of trees within each model grown on different subsets of the full training set prevents over-fitting (which is commonly seen in single decision trees) and reduces variance. Finally, in empirical comparisons of many machine learning methods, RFs consistently rank among the best approaches [43], and we too found RFs to produce the best outputs among the classifiers tested. We employ RFs for both cross-study and intra-study validations using different testing and training datasets (details in Methods).

Before analysis, several data pre-processing steps are required to remove time as an implicit feature in our physiological datasets. First, data is normalized and aggregated subject-by-subject to eliminate short-term fluctuations and daily diurnal rhythms. From these normalized datasets, mean and quantiles are calculated for adjacent 30 minute time windows (see Figure 3); these first- and second-order statistical measures are the features provided to the machine learning algorithm. Two RF models are trained to detect the post-exposure class at distinct time epochs: one model is tuned to detect subtle data markers during the incubation phase prior to fever, while the second model

is tuned for the early prodromal phase (i.e., onset of overt febrile symptoms) where temperature-related features emerge as powerful discriminants. The training data for the pre-exposure class for both models is a subset of baseline data prior to challenge and the quantity of training data has been balanced for the negative (pre-exposure) and positive (post-exposure) classes to avoid biasing one class over the other. For the rest of our analysis, data from 12h before and 24h after challenge are excluded from performance metrics due to differences in animal handling and sedation for exposure. Additional details on data pre-processing and algorithm development may be found in the Methods section.

One output of RFs is a measure of relative feature importance; that is, which features provide the most accurate separation between classes. The most discriminating features for the pre-fever and post-fever RF models vary among four feature types derived from temperature, ECG, blood pressure, and respiration measurements. (See Supplementary Table 1 and Supplementary Figures 2-3 for a complete listing of most discriminating features.) The algorithm reports features that follow clinical symptomology, namely that core temperature in the post-fever, prodromal model is the highest ranking in feature importance. Before fever, however, subtle ECG and blood pressure derived features seem to be the highest ranking in feature importance, as is observed at the earliest stages of sepsis [28-31] (further noted in the Discussion below).

Intra-study (3-fold) cross-validations

Intra-study tests (i.e., training and tested with data from the same NHP study) are used first for testing our model's accuracy and early warning capability. RFs are built using two-thirds of the subjects then tested for each subject in the remaining one-third of left-out subjects; the process is repeated three times, allowing each subject in the study to be evaluated once as part of a left-out test data set. The resulting outcomes for all subjects are combined and evaluated in subsequent performance metrics. Three-fold cross-validation is chosen due to limited study sample sizes. These intra-study tests necessarily hold constant factors such as animal species, exposure route, and virus, and thus provide insight into the model's performance when these factors are known or are

constant. Figure 4 shows representative examples of our algorithm's output for each intra-study test. Every 30 minutes, the combined score (see Figure 2) of the pre- and post-fever forests is plotted, representing the *a posteriori* probability that the subject is in the "exposed" class. In other words, values closer to 1.0 indicate a higher confidence prediction for a subject having been exposed to the virus. Qualitatively, we note that most subjects' scores rise around the challenge time (though data 12h before and 24h after exposure are disregarded). To quantify performance, we calculate probability of infection detection P_d and probability of false declaration P_{fa} for the collection of system outputs (updated every 30 minutes). Associated with these are the 95% confidence intervals for a standard Gaussian. In cases where no false declarations were made within the study sample, we provide an upper bound on P_{fa}. For the MARV IM study (system P_d=0.95±0.008, P_{fa} <0.002, Δt_{mean}=74.5±6.0h), the scores rise sharply after challenge and remain high throughout the remainder of the study. The MARV aerosol (system P_d=0.79±0.02, P_{fa} <0.002 Δt_{mean}=44.4±26.1h) and EBOV aerosol (system P_d=0.65±0.02, P_{fa}=0.01±0.005, Δt_{mean}=23.0±30.3h) studies show moderate elevations at challenge time and fluctuate the first few days before rising sharply 12 to 24h before acute fever (vertical red line). This behavior can be explained by trends in the individual forest scores. The pre-fever forest is trained on data with subtle, sub-clinical changes from pre-exposure baseline which become more obvious and detectable in the hours leading up to fever onset (and when the animal is anesthetized prior to challenge). Variability in the combined score before fever can be understood both by considering the individual animal's immune response to the pathogen, and the inter-individual variability of this response when training the algorithm across subjects. Furthermore, variability in the pre-fever results and lower early warning time for the EBOV study may be due to a much lower target exposure dose (100pfu target) than either of the MARV studies (1000pfu target). After febrile symptoms, the post-fever forest dominates the score as it indicates a strong and easily detectable deviation from the baseline and is how current clinical diagnosis is largely based.

To quantitatively assess whether a subject has been exposed, we use a false positive threshold method (details in Methods section) to build a binary decision from the RF models, then employed a binary integration step to make a final declaration that a subject is exposed. These two steps afford much greater sensitivity and specificity than

relying on RF model score outputs alone [44]. Briefly, using baseline pre-exposure data, we threshold scores from each RF model and make an 'initial detection' decision every 30 minutes. Next, we perform binary integration which accumulates the number of positive detections, m, observed in the past n time steps. At each time step, if the accumulated detections are greater than or equal to m (here we used m=11 and n=24), we output a 'declaration' that the subject is in the exposed class at that time step. We find threshold values for each RF model by sweeping across a series of possible thresholds from 0 to 1. For each threshold in the series, the proportion of false declarations P_{fa} is calculated using 3-fold cross validation, in the same manner as described for the RFs. Thresholds are estimated as the smallest value for each RF model that supported a desired constant false positive level (here we chose P_{fa}=0.01). Figure 4 shows our algorithm's combined score output (after thresholding and binary integration, see Figure 2b), declarations (green triangles), and onset of fever (red vertical lines) for three representative subjects in each study. The time between our algorithm's first true declaration (green line) and fever onset is defined here as the early warning time (Δt). (Each subject's early warning times, as well as additional algorithm performance parameters, may be found in Supplementary Table 2.) As with P_d and P_{fa}, we report 95% confidence intervals associated with each Δt. However, unlike P_d and P_{fa}, the number of trials for early warning time are small (20 subjects per test at most) so the confidence intervals are based on t-distributions with the degrees of freedom equal to the number of subjects minus 1.

The early warning capacity for these intra-study tests demonstrate the ability to find meaningful Δt values when the animal species, exposure route, and viral agent is known. We can imagine a context where such information is known, such as a healthcare worker experiencing an accidental needle stick in a known outbreak, or a laboratory employee after an accidental protective equipment breech. However, most exposures will occur when many of these variables are unknown or impossible to know, which emphasizes the need to experiment with testing and training our algorithm across these variables.

Cross-study Validations

Next, we used cross-study validations to indicate our model's extensibility beyond a given animal model, pathogen, or exposure route. In one version of a cross-study validation, all data from one NHP study are used to train RF models, all data from a second study are used to test that model, and an identical false positive thresholding and binary integration method for detection/declaration as used above is applied. Algorithm outputs and detection plots are interpreted identically as in the intra-study validation tests. Figure 5 shows one representative subject's output for each of the (train/test) MARV intramuscular/MARV aerosol (system P_d=0.0.81±0.02, P_{fa} =0.04±0.01, Δt_{mean}=42.5±22.1h) and EBOV aerosol / MARV aerosol (system P_d=0.72±0.02, P_{fa} =0.01±0.005, Δt_{mean}=28.3±16.2h) cross-study validations. These combinations are chosen to hold the pathogen and exposure routes constant, respectively. The EBOV aerosol / MARV aerosol validation test also uses studies with different target dose exposure levels, which may explain the lower early warning time; despite this, we still observe nearly 1 day of early warning.

In another version of a cross-study validation, we tested the most generalized scenario where all data across all three studies are used to test and train a RF model. In this aggregate study where the species of animal, exposure route, virus, nor target dose are held constant, we find a system P_d=0.80±0.01, P_{fa} <0.0005, and Δt_{mean}=52.8±12.9h. These results strongly suggest that our model is not limited to particular viruses or exposure routes, but rather is capable of indicating a general patho-physiological state during the viral incubation period in NHPs.

Evaluating Algorithm Performance

We evaluated our algorithm's performance by analyzing the probability of detection (P_d, i.e., correctly declaring a subject as being exposed after the viral challenge) versus false positives (P_{fa}, i.e., incorrectly declaring a subject as exposed before the viral challenge), known as a receiver operating characteristic (ROC) curve [45]. ROC curves describe the sensitivity (P_d) and specificity (1-P_{fa}, i.e., not informative of the causative

agent) of a test and can be partially summarized by the area under the curve (AUC). An AUC of 1.0 refers to a perfectly sensitive and specific detector, whereas a value of 0.5 indicates that the test cannot distinguish between classes and is no better than a coin-flip. Figure 6 shows ROC curves for the MARV aerosol intra-study, the MARV IM/MARV aerosol cross-study tests, and the aggregate study test using all available data; additional ROC curves for all intra-study and cross-study validations can be found in the Supplementary Figures 4-5. We conclude that for each intra-study and cross-study test that the pre-fever AUC is ≥0.90, and thus each pre-fever RF model has significant discriminating power for early detection (details in Supplementary Table 3). All post-fever RF models have AUC values approaching one, indicating nearly perfect performance during febrile symptoms as may be expected given such as clear anomaly compared to baseline values.

Perhaps the most clinically useful metric of our algorithm is the early warning time, defined as the time difference between our algorithm's first correct 'declaration' and the onset of fever (1.5ºC above a diurnal baseline [38] sustained for two hours). Another useful metric from an algorithm development perspective is the ROC AUC for different subsets of study data collected before fever (e.g. the interval where early warning is meaningful). This pre-fever ROC AUC provides a robust metric for performance comparisons both across studies and evaluating system design trade-offs such as reduced feature sets, as discussed below.

Extending to non-invasive monitoring platforms

Physiological data features provided to our algorithm were collected using surgically implanted monitoring devices; such data could never be expected from military service members, health care workers responding to an outbreak, hospital patients, or the general public. As an *in silico* simulation for limiting our dataset to what may be collected using a wearable-type monitoring device, we reduced the considered feature set to include only certain subsets: ECG-only, ECG and temperature, heart rate and temperature, temperature alone, and heart rate alone. Successful use of ECG data as a predictor of physiological compensatory potential during shock has been reported

[28,29]. Ambulatory Holter monitor devices collect exactly this type of data [46], as do even less obtrusive devices for performance athletes. Figure 7 shows algorithm output for one representative subject and ROC curves for the MARV aerosol study using this ECG-only feature subset (including RR, QRS, PR, and QT intervals; the relative importance of each feature is shown in Supplementary Figure 6). Although the sensitivity of this ECG-only algorithm decreases slightly relative to the baseline feature set (with a P_d=0.78±0.02 at a P_{fa}<0.001 vs a P_d=0.79±0.02 at the same P_{fa}), the mean early warning time of 51.1±23.0h is still very clinically useful. Results for other reduced feature subsets of the MARV aerosol study are provided in Figure 8 and additional feature importance metrics and corresponding ROC curves may be found in Supplementary Figures 6-7.

Discussion

Non-biochemical detection of viral incubation periods using only physiological data presents a fundamentally new approach to infectious disease care. Previous work has shown that reducing transmission during the viral incubation period is as or more effective an intervention as reducing the inherent transmissibility (R_0) of the pathogen in controlling emerging outbreaks [9]. However, there is no existing method to detect this pre-symptomatic incubation period extensible to mobile settings or wearable sensor systems. We present the first attempt to build a multi-modal machine learning algorithm capable of determining this incubation period using physiological signals of NHPs infected with viral hemorrhagic fevers. Using the Random Forest method we avoid building over-fit models, and successful testing and training on different subsets of data demonstrate that we avoid over-fitting. Further, cross-study validations show the promise of extending this approach beyond a given animal model, exposure method, or virus. All intra-study and cross-study validations resulted in positive mean early warning times, with times that would be actionable (>20h) for intervention or other preventive measures. While we chose a target system P_{fa}~0.01 that was supported by the limited subject numbers in the studies available, this would not lead to an acceptable daily false alarm rate – reducing this critical system parameter to more

clinically-acceptable levels (we estimate P_{fa} ~10^{-3} or less) is the subject of on-going work, and may require larger sample sizes or more refined processing algorithms.

We postulate that immuno-biological events – particularly systemic release of pro-inflammatory chemokines and cytokines from infected phagocytes [47-51], as well as afferent signaling to the central nervous system [52,53] – are recapitulated in hemodynamic, thermoregulatory, or cardiac signals which may be more easily measured and assessed than biomolecule markers for viral infection (via sequencing [23,24,26] or immunocapture approaches [15,16]). For instance, prostaglandins (PG) are up-regulated upon infection (including EBOV [54,55]) and intricately involved in the non-specific "sickness syndrome" [56]; the PGs are also known to be potent vascular mediators [57] and endogenous pyrogens [58,59]. Past work has clarified how tightly integrated, complex, and oscillating biological systems can become uncoupled [60-62] during trauma [63] or critical illness [31,64] which would be captured in the comprehensive, multi-modal physiological datasets used in our present study. Rigorously pursuing this hypothesis would require additional high temporal resolution datasets, including high-resolution biochemical, immunological, neurological, and cardiovascular information.

Previous work on genomic [23,24] profiles of peripheral blood cells following acute influenza infection indicate specific host responses at just ~45h following exposure, corresponding to ~35h of early warning time. Our combined results suggest that the classic understanding of an asymptomatic incubation phase may be incomplete: during viral incubation, subtle sub-clinical cues (both genomic, transcriptional, and physiological) can be detectable with sufficiently high-resolution sensor and analysis systems. Better understanding of how biomolecular changes are captured in systemic physiological signals during viral infection would open further opportunities for better therapeutic administration both before and during infection, quarantine or isolation, and vaccine development.

Detecting pathogen exposure before self-reporting or overt clinical symptoms affords great opportunities in clinical care and public health measures. However, given the consequences of using some of these interventions and the lack of etiological agent

specificity in our algorithm, we envision our current approach to be a trigger for 'low-regret' actions rather than necessarily guiding medical care. For instance, using our high sensitivity approach as an alert for limited high specificity confirmatory diagnostics (such as sequencing or PCR-based) could lead to considerable cost savings (an "alert-confirm" system). Public health response following a bioterrorism incident could also benefit from triaging those exposed from the "worried well." Ongoing work focuses on adding enough causative agent specificity to discern between bacterial and viral pathogens; even this binary classification would be of use for front-line therapeutic or mass casualty uses. Eventually, we envision a system that could give real-time prognostic information, even before obvious illness, guiding patients and clinicians in diagnostic or therapeutic use with better time resolution than ever before.

Methods

Viruses

The Marburg Angola isolate used was USAMRIID challenge stock "R17214" (Marburg virus H.sapiens-tc/ANG/2005/Angola-1379c). Cynomolgus macaques were exposed to Ebola virus/H.sapiens-tc/COD/1995/Kikwit-9510621 (EBOV) at a target dose of 100 pfu (7U EBOV; USAMRIID challenge stock "R4415"; GenBank # KT762962).

Description of Studies

Dr. William Pratt provided physiological data in NSS format (Notocord, Inc.) from studies previously conducted at the United States Army Medical Research Institute of Infectious Diseases (USAMRIID). Research was conducted under an IACUC approved protocol in compliance with the Animal Welfare Act, PHS Policy, and other Federal statutes and regulations relating to animals and experiments involving animals. The facility where this research was conducted is accredited by the Association for Assessment and Accreditation of Laboratory Animal Care, International and adheres to principles stated in the Guide for the Care and Use of Laboratory Animals, National Research Council, 2011. In each study, remote telemetry devices (Konigsberg Instruments, Inc., T27F for MARV studies and T37F for 3 subjects in the EBOV study, and Data Sciences International Inc., L11 for 3 subjects in the EBOV) were implanted 3 to 5 months before exposure, and, if used, a central venous catheter was implanted 2 to 4 weeks before. NHPs were transferred into BSL4 containment 5 to 7 days before viral exposure, and baseline pre-exposed data collected for 4 to 6 days before. Subjects were exposed under sedation via either aerosol or intramuscular injection depending on the study. The exposure time used in our model is based upon the time of intramuscular injection or when a subject was returned to the cage following aerosol exposure (~20 min). All subjects were monitored until death or the completion of the study. The devices measure several raw physiological signals, which were translated to blood pressure (sampling frequency f_s = 250Hz), ECG (f_s = 500Hz), temperature (f_s = 50Hz), and pulmonary (f_s = 50Hz) features. We analyzed data from three separate studies, detailed in Table 1.

Physiological Data Pre-processing

Physiological data is time dependent (that is, sequential data) and is subject to short-term fluctuations and daily diurnal rhythms. RF classifiers, however, require time and subject independent data. To reduce diurnal and subject dependencies from the data, each subject is pre-processed individually. The first step is to estimate baseline diurnal statistics of the data by computing a mean, μ_i, and standard deviation, σ_i , for 30-minute intervals $i = 1, \ldots, 48$ over an average 24-hour pre-exposure period. The data for that time of day is normalized by subtracting the mean and dividing by the standard deviation, $(x_i(j) - \mu_i)/\sigma_i$ for each data sample *j* in the i^{th} interval. Data are then partitioned into sequential *k*-minute blocks and aggregated by calculating a set of three summary statistics on each block: mean and 25% and 75% quantiles. These summary statistics calculated on each time-independent signal are the input features for the random forest algorithm. For example, 30-minute blocks for two days of 4 raw physiological signals yields 96 time points with 12 data features. Although the normalization period and aggregation blocks (*k*) are not required to be the same, we have chosen a common interval of 30 minutes for both. Data samples that correspond to measurements before challenge are labeled "0" to denote the pre-exposed class and those after challenge are labeled "1" to denote the post-exposure class.

Random Forest Algorithm

Our model is composed of two random forests (RFs): one RF is grown using training data prior to fever onset and an equal number of randomly chosen negative data samples from the pre-exposure class. Since the number of subjects in each study is very small, we do not have a separate validation set. However, test data is always held out until the final evaluation step. Each RF contains 50 classification decision trees grown on random subsets of data and features. The trees cast their "votes" for class "0" or "1", and the forest returns the class with the most votes. This process helps prevent overfitting, which single decision trees tend to do. RFs are particularly good for calculating feature importance metrics, and we use these metrics to find the most predictive features for hard to classify (pre-fever) days. Initially all features are considered, but once the subset of most predictive features is determined within a cross-

validation training set, all RF's are regrown (same training set) on this 15 feature subset to produce the final models upon which the corresponding cross-validation testing set performance results are based. Relative importance scores for each of the top 15 features from each study are provided in the Supplementary Materials.

Model Performance Evaluation: Cross-Study and Inter-study validations

Model performance may be evaluated by separating subjects into testing and training sets. We conduct two modes of evaluation: cross-study, where testing and training data are from different studies (and thus can vary in subject species, virus, and exposure route), and intra-study, where both testing and training datasets are from the same study (with constant subject species, pathogen, and exposure routes) thus allowing model evaluation across individuals. We used a 3-fold cross-validation for the intra-study tests by randomly assigning subjects into three partitions. Subjects from two of those partitions form the training set to build the model, while one subject at a time from the held-out partition is tested against that model. Model building and subject testing is repeated for all subjects in a study. Most cross-study evaluations used all data from one study to train the model, and all subjects of another study are tested using that model. In the aggregated cross-study validation, we used a 3-fold cross-validation just as with the intra-study tests, including random assignment of subjects into the three partitions. Each partition included subjects from each of the three studies.

False Positive Thresholding, Binary Integration and Algorithm Performance Metrics

We make declarations of exposure using a two-stage detection process (see Figure 2). In stage one of the detection process, RF model prediction scores (between 0 and 1 for every 30 minute interval) are thresholded (i.e., a value of 1 is returned if the RF model score is greater than or equal to the threshold found above) to form a series of initial detections for the model every 30 minutes.

These initial detections from each RF model are subjected to a second-stage detection test to further reduce the false alarm rate. During the second stage, binary integration is performed over a sliding window of the past n initial detections. The accumulated

detections are normalized by n, giving a mean score for the pre- and post-fever RF models. Next, scores are combined by taking the maximum of the pre- or post-fever values to create a single time series. At each 30 minute time interval, this combined score is compared to a final declaration threshold of m/n, where $m \leq$ (we selected n=24 for a system latency of no more than 12 hours and selected m=11 which approximates the optimum binary integration threshold for a steady signal in noise [65]; performance is relatively insensitive to small deviations in m or n). The algorithm makes a 'declaration' that the subject is in the exposed class when the combined score is greater than or equal to m/n; if the threshold is not met, the algorithm assigns the subject to the 'not exposed' class for that time epoch. Note that n samples are required before a declaration can be made, so following the start of data collection or the end of an exclusion period (the 24h period following the challenge), no declarations are reported in the first 30n minutes (for n=24, this accumulation period effectively extends the exclusion period to 36 hours post-challenge).

Threshold levels for the pre- and post-fever RFs are estimated by analyzing false alarm rates (Type I errors) of the final declarations versus threshold levels (swept from 0 to 1). We define the probability of false alarm (or P_{fa}) as

$$P_{fa} = \frac{\#\ False\ Positives}{\#\ True\ Negatives + \#\ False\ Positives}$$

To enforce a desired significance level (we choose $P_{fa} = 0.01$), we evaluate P_{fa} for the final declarations for subjects in the current partition and estimate the smallest threshold needed in the stage-one detection shown in Figure 2b. This approach is repeated for three partitions in each study, resulting in independent estimates of the threshold pair (pre- and post-fever) for each partition. While the desired $P_{fa} = 0.01$, the final overall system P_{fa} may be higher or lower.

To evaluate system-level performance, we define probability of correct declaration P_d as:

$$P_d = \frac{\#\ True\ Positives}{\#\ True\ Positives + \#\ False\ Negatives}$$

and P_{fa} as above, where the True Positives, False Positives, True Negatives and False Negatives are evaluated on the final declaration outputs of Figure 2b. When reporting P_d and P_{fa} for a study, we include the 95% confidence interval based on standard normal distributions since the number of trials per study is large (>2000). Although some correlation is likely within a binary integration window of 30n minutes, we assume independence for trials separated by at least 30n minutes. We generate receiver operating characteristic (ROC) curves to measure system performance by calculating P_d vs P_{fa} at a series of threshold values (sweeping the first-stage detection threshold but holding the second-stage m/n threshold constant) and quantify the system performance with the ROC area under the curve (AUC), where an AUC=1.0 indicates perfect performance and AUC=0.5 indicates that the model is no better than a coin toss. Sensitivity (P_d) is expected to be highest after febrile symptoms are apparent. To distinguish the sensitivity of the system during the pre- and post-fever epochs, P_d is calculated independently for subsets of positive data that occur before and after the onset of fever. The result is two ROC curves and corresponding AUCs: one evaluated on positive data restricted to pre-fever time samples and the other restricted to post-fever time samples. The negative data and two-stage detection process are identical for both ROC curves.

In a clinically or military-deployed early-warning system, it may be desirable to calculate P_d and P_{fa} on a per-device or per-day basis. However, for this proof-of-concept study, the limited pool of subjects available (N=20 total) necessitates calculating P_d and P_{fa} across all 30-minute test points that are not in the exclusion window (12h before and 24h after exposure). This approach includes false negatives that may occur after an initial early-warning declaration is made, and thus provides a conservative estimate of the device sensitivity which we predict will further increase with larger sample sizes and more refined processing algorithms.

Another important measure of system performance is the mean early warning time. The early warning time for an individual subject is defined as the time of the first true declaration (excluding data from the 24 h interval immediately following the challenge) minus the time of fever onset (defined as 1.5ºC above a diurnal baseline [38] sustained

for two hours). Early warning times vary across subjects in a study, so the mean value is calculated across all subjects to characterize the early warning time afforded by the system.

Acknowledgements

This work is sponsored by the Department of the Army and Defense Threat Reduction Agency under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government or reflect the views or policies of the US Department of Health and Human Services. We thank Jason Williams for his excellent graphical support, Dr. Brian Telfer for his thoughtful manuscript comments, and Amanda Casale for her expert statistics guidance.

Competing Financial Interests

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at nature.com.

Provisional U.S patent application 62/193,961 was filed July 2015.

Author Contributions

WP, GC and AS conceived the study. LM, SD, JF, TP, MH, CC, AR, and AS processed and analyzed the physiology data and built the algorithm. AH, LH, AG, JT, FR, and WP collected the physiology data and supervised the animal studies. LM, TP, SD, and AS wrote the manuscript. AS supervised the study.

References

1. Liu R, Wang X, Aihara K, Chen L (2014) Early Diagnosis of Complex Diseases by Molecular Biomarkers, Network Biomarkers, and Dynamical Network Biomarkers. Medicinal Research Reviews 34: 455-478.
2. Bociaga-Jasik M, Piatek A, Garlicki A (2014) Ebola virus disease - pathogenesis, clinical presentation and management. Folia Med Cracov 54: 49-55.
3. Tosh PK, Sampathkumar P (2014) What Clinicians Should Know About the 2014 Ebola Outbreak. Mayo Clinic Proceedings 89: 1710-1717.
4. Bausch DG, Hadi CM, Khan SH, Lertora JJL (2010) Review of the Literature and Proposed Guidelines for the Use of Oral Ribavirin as Postexposure Prophylaxis for Lassa Fever. Clinical Infectious Diseases 51: 1435-1441.
5. Stiver G (2003) The treatment of influenza with antiviral drugs. Cmaj 168: 49-56.
6. Khan AS, Tshioko FK, Heymann DL, Le Guenno B, Nabeth P, et al. (1999) The Reemergence of Ebola Hemorrhagic Fever, Democratic Republic of the Congo, 1995. Journal of Infectious Diseases 179: S76-S86.
7. Pandey A, Atkins KE, Medlock J, Wenzel N, Townsend JP, et al. (2014) Strategies for containing Ebola in West Africa. Science 346: 991-995.
8. Eichner M (2003) Case isolation and contact tracing can prevent the spread of smallpox. Am J Epidemiol 158: 118-128.
9. Fraser C, Riley S, Anderson RM, Ferguson NM (2004) Factors that make an infectious disease outbreak controllable. Proceedings of the National Academy of Sciences of the United States of America 101: 6146-6151.
10. Evans AS, Kaslow RA (1997) Viral infections of humans : epidemiology and control: New York : Plenum Medical Book Co. xxxviii, 1078 p., 1071 p. of plate p.
11. American Public Health Association. (1995) Control of communicable diseases manual : an official report of the American Public Health Association. Washington, DC: American Public Health Association. pp. v.
12. Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE, et al. (2009) Incubation periods of acute respiratory viral infections: a systematic review. Lancet Infect Dis 9: 291-300.
13. Eichner M, Dowell SF, Firese N (2011) Incubation Period of Ebola Hemorrhagic Virus Subtype Zaire. Osong Public Health and Research Perspectives 2: 3-7.
14. Pavlin BI (2014) Calculation of incubation period and serial interval from multiple outbreaks of Marburg virus disease. BMC Res Notes 7: 906.
15. Bausch DG, Rollin PE, Demby AH, Coulibaly M, Kanu J, et al. (2000) Diagnosis and Clinical Virology of Lassa Fever as Evaluated by Enzyme-Linked Immunosorbent Assay, Indirect Fluorescent-Antibody Test, and Virus Isolation. Journal of Clinical Microbiology 38: 2670-2677.
16. Drosten C, Kümmerer BM, Schmitz H, Günther S (2003) Molecular diagnostics of viral hemorrhagic fevers. Antiviral Research 57: 61-87.
17. Sedlak RH, Jerome KR (2013) Viral diagnostics in the era of digital polymerase chain reaction. Diagnostic Microbiology and Infectious Disease 75: 1-4.
18. Muldrew KL (2009) Molecular diagnostics of infectious diseases. Current Opinion in Pediatrics 21: 102-111.
19. Mahony JB (2008) Detection of Respiratory Viruses by Molecular Methods. Clinical Microbiology Reviews 21: 716-747.

20. Kortepeter MG, Bausch DG, Bray M (2011) Basic clinical and laboratory features of filoviral hemorrhagic fever. J Infect Dis 204 Suppl 3: S810-816.
21. Ksiazek TG, Rollin PE, Williams AJ, Bressler DS, Martin ML, et al. (1999) Clinical virology of Ebola hemorrhagic fever (EHF): Virus, virus antigen, and IgG and IgM antibody findings among EHF patients in Kikwit, Democratic Republic of the Congo, 1995. Journal of Infectious Diseases 179: S177-S187.
22. Drosten C, Gottig S, Schilling S, Asper M, Panning M, et al. (2002) Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, dengue virus, and yellow fever virus by real-time reverse transcription-PCR. J Clin Microbiol 40: 2323-2330.
23. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, et al. (2009) Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infection in Humans. Cell host & microbe 6: 207-217.
24. Woods CW, McClain MT, Chen M, Zaas AK, Nicholson BP, et al. (2013) A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. PLoS One 8: e52198.
25. Caballero IS, Yen JY, Hensley LE, Honko AN, Goff AJ, et al. (2014) Lassa and Marburg viruses elicit distinct host transcriptional responses early after infection. BMC Genomics 15: 960.
26. Shurtleff AC, Whitehouse CA, Ward MD, Cazares LH, Bavari S (2015) Pre-symptomatic diagnosis and treatment of filovirus diseases. Frontiers in Microbiology 6: 108.
27. Scheff JD, Mavroudis PD, Calvano SE, Androulakis IP (2013) Translational applications of evaluating physiologic variability in human endotoxemia. Journal of clinical monitoring and computing 27: 405-415.
28. Korach M, Sharshar T, Jarrin I, Fouillot JP, Raphael JC, et al. (2001) Cardiac variability in critically ill adults: influence of sepsis. Crit Care Med 29: 1380-1385.
29. Chen W-L, Kuo C-D (2007) Characteristics of Heart Rate Variability Can Predict Impending Septic Shock in Emergency Department Patients with Sepsis. Academic Emergency Medicine 14: 392-397.
30. Ahmad S, Ramsay T, Huebsch L, Flanagan S, McDiarmid S, et al. (2009) Continuous multi-parameter heart rate variability analysis heralds onset of sepsis in adults. PLoS One 4: e6642.
31. Scheff JD, Mavroudis PD, Foteinou PT, Calvano SE, Androulakis IP (2012) Modeling Physiologic Variability in Human Endotoxemia. Critical reviews in biomedical engineering 40: 313-322.
32. Papaioannou VE, Chouvarda IG, Maglaveras NK, Pneumatikos IA (2012) Temperature variability analysis using wavelets and multiscale entropy in patients with systemic inflammatory response syndrome, sepsis, and septic shock. Crit Care 16: R51.
33. Williamson ED, Savage VL, Lingard B, Russell P, Scott EA (2007) A biocompatible microdevice for core body temperature monitoring in the early diagnosis of infectious disease. Biomed Microdevices 9: 51-60.
34. Madan A, Cebrian M, Lazer D, Pentland A (2010) Social sensing for epidemiological behavior change. Proceedings of the 12th ACM international conference on Ubiquitous computing. Copenhagen, Denmark: ACM. pp. 291-300.
35. Sun G, Abe N, Sugiyama Y, Nguyen QV, Nozaki K, et al. (2013) Development of an infection screening system for entry inspection at airport quarantine stations

using ear temperature, heart and respiration rates. Conf Proc IEEE Eng Med Biol Soc 2013: 6716-6719.
36. Pantelopoulos A, Bourbakis NG (2010) A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 40: 1-12.
37. Banaee H, Ahmed MU, Loutfi A (2013) Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges. Sensors (Basel, Switzerland) 13: 17472-17500.
38. Laupland KB (2009) Fever in the critically ill medical patient. Crit Care Med 37: S273-278.
39. Kojic D, Siegler BH, Uhle F, Lichtenstern C, Nawroth PP, et al. (2015) Are there new approaches for diagnosis, therapy guidance and outcome prediction of sepsis? World J Exp Med 5: 50-63.
40. Breiman L (2001) Random forests. Machine Learning 45: 5-32.
41. Rish I (2001) An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence 3: 41-46.
42. Cover T, Hart P (1967) Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13: 21-27.
43. Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning. Pittsburgh, Pennsylvania, USA: ACM. pp. 161-168.
44. Dillard GM (1967) A moving-window detector for binary integration. Information Theory, IEEE Transactions on 13: 2-6.
45. Zou KH, O'Malley AJ, Mauri L (2007) Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation 115: 654-657.
46. Enseleit F, Duru F (2006) Long-term continuous external electrocardiographic recording: a review. Europace: European Pacing, Arrhythmias, And Cardiac Electrophysiology: Journal Of The Working Groups On Cardiac Pacing, Arrhythmias, And Cardiac Cellular Electrophysiology Of The European Society Of Cardiology 8: 255-266.
47. Martinez FO, Sica A, Mantovani A, Locati M (2008) Macrophage activation and polarization. Frontiers in Bioscience 13: 453-461.
48. Hayden FG, Fritz R, Lobo MC, Alvord W, Strober W, et al. (1998) Local and systemic cytokine responses during experimental human influenza A virus infection. Relation to symptom formation and host defense. J Clin Invest 101: 643-649.
49. Leroy EM, Baize S, Volchkov VE, Fisher-Hoch SP, Georges-Courbot MC, et al. (2000) Human asymptomatic Ebola infection and strong inflammatory response. The Lancet 355: 2210-2215.
50. Gupta M, Mahanty S, Ahmed R, Rollin PE (2001) Monocyte-Derived Human Macrophages and Peripheral Blood Mononuclear Cells Infected with Ebola Virus Secrete MIP-1 α and TNF- α and Inhibit Poly-IC-Induced IFN- α in Vitro. Virology 284: 20-25.
51. Hensley LE, Young HA, Jahrling PB, Geisbert TW (2002) Proinflammatory response during Ebola virus infection of primate models: possible involvement of the tumor necrosis factor receptor superfamily. Immunology Letters 80: 169-179.
52. Tracey KJ (2002) The inflammatory reflex. Nature 420: 853-859.
53. Beishuizen A, Thijs LG (2003) Endotoxin and the hypothalamo-pituitary-adrenal (HPA) axis. J Endotoxin Res 9: 3-24.

54. Geisbert TW, Young HA, Jahrling PB, Davis KJ, Larsen T, et al. (2003) Pathogenesis of Ebola hemorrhagic fever in primate models: evidence that hemorrhage is not a direct effect of virus-induced cytolysis of endothelial cells. Am J Pathol 163: 2371-2382.
55. Wahl-Jensen V, Kurz S, Feldmann F, Buehler LK, Kindrachuk J, et al. (2011) Ebola virion attachment and entry into human macrophages profoundly effects early cellular gene expression. PLoS Negl Trop Dis 5: e1359.
56. Saper CB, Romanovsky AA, Scammell TE (2012) Neural circuitry engaged by prostaglandins during the sickness syndrome. Nat Neurosci 15: 1088-1095.
57. Funk CD (2001) Prostaglandins and Leukotrienes: Advances in Eicosanoid Biology. Science 294: 1871-1875.
58. Sugimoto Y, Narumiya S, Ichikawa A (2000) Distribution and function of prostanoid receptors: studies from knockout mice. Progress in Lipid Research 39: 289-314.
59. Ek M, Engblom D, Saha S, Blomqvist A, Jakobsson P-J, et al. (2001) Inflammatory response: Pathway across the blood-brain barrier. Nature 410: 430-431.
60. Godin PJ, Buchman TG (1996) Uncoupling of biological oscillators: a complementary hypothesis concerning the pathogenesis of multiple organ dysfunction syndrome. Crit Care Med 24: 1107-1116.
61. Goldberger AL, Peng CK, Lipsitz LA (2002) What is physiologic complexity and how does it change with aging and disease? Neurobiol Aging 23: 23-26.
62. Bravi A, Longtin A, Seely AJE (2011) Review and classification of variability analysis techniques with clinical applications. BioMedical Engineering OnLine 10: 90-90.
63. Cancio LC, Batchinsky AI, Baker WL, Necsoiu C, Salinas J, et al. (2013) Combat casualties undergoing lifesaving interventions have decreased heart rate complexity at multiple time scales. J Crit Care 28: 1093-1098.
64. Scheff JD, Calvano SE, Androulakis IP (2013) Predicting critical transitions in a model of systemic inflammation. J Theor Biol 338: 9-15.
65. Shnidman DA (1998) Binary integration for Swerling target fluctuations. Ieee Transactions on Aerospace and Electronic Systems 34: 1043-1053.

Virus	Exposure method	Subjects	Species	Monitoring system	Target dose (pfu)
EBOV	Aerosol	6	Cynomolgus	3 subjects with ITS T37F 3 subjects with DSI L11	100
MARV	Aerosol	5	Rhesus	ITS T27F	1000
MARV	IM	9	Cynomolgus	ITS T27F	1000

Table 1: Summary of NHP studies used. The EBOV study compared two different physiological monitoring systems but data was combined and treated identically.

Figure 1

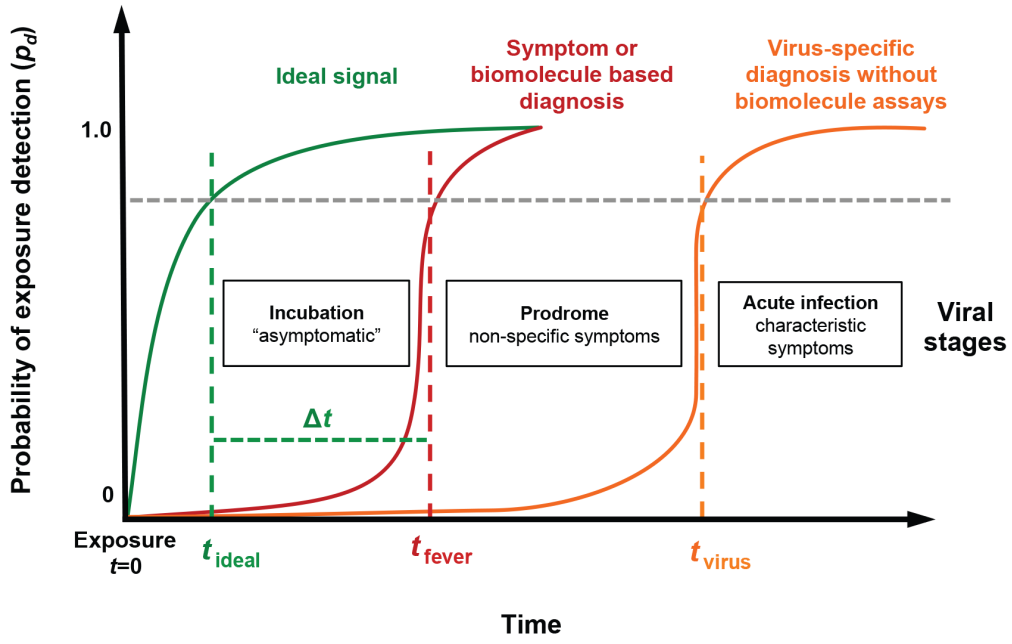

Figure 1: Notional schematic of the probability of detection (P_d) for current symptoms-based detection (red curve) and an ideal signal (green curve) versus time (viral exposure at t=0), overlaid with a typical evolution of symptoms. An ideal sensor and analysis system would be capable of detecting exposure for a given P_d (and probability of false alarm, P_{fa}) during the incubation period (t_{ideal}), well before the non-specific symptoms of the prodrome (t_{fever}). We define the difference $\Delta t = t_{fever} - t_{ideal}$ as the *early warning time* (details below).

Figure 2

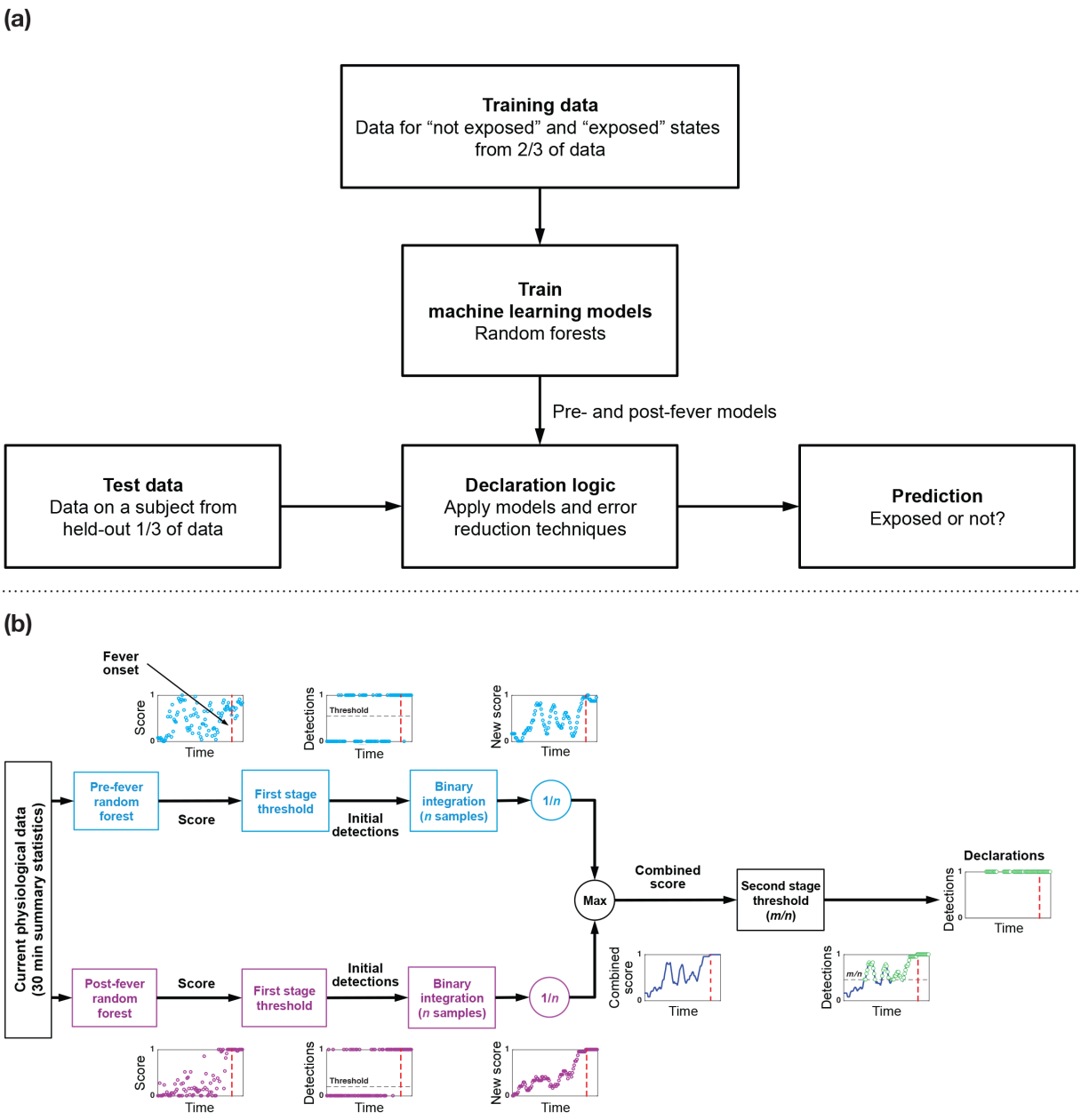

Figure 2: Workflow of our (a) classification approach using random forests and (b) block diagram of a two-stage detection algorithm to reduce false alarms. The detection scheme comprises two distinct stages: after the random forest model score output, an a priori determined threshold (based on a desired P_{fa}) is applied to yield initial detections. These are then subjected to a binary integration step of the past n samples, and the maximum value of the pre- and post-fever models are taken to produce a single time series. A second stage m of n detection is applied, which finally produced a final 'declaration' of being exposed or not. See Methods for detailed descriptions.

Figure 3

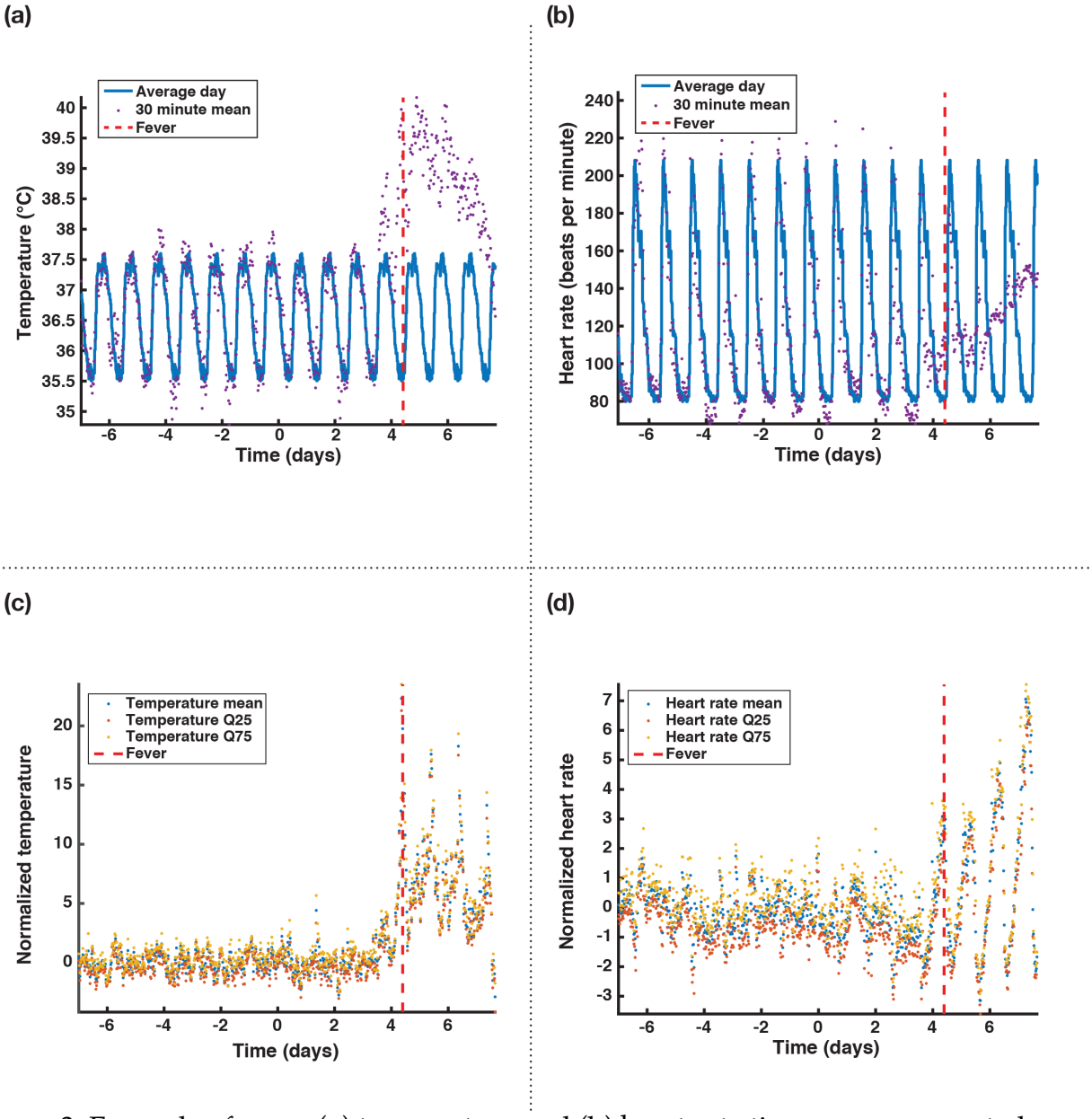

Figure 3: Example of mean (a) temperature and (b) heart rate time courses reported every 30 min from on subject in the MARV aerosol study. The blue curves indicate an average diurnal value for this subject before exposure. Same (c) temperature and (d) heart rate data after normalization and calculation of mean, standard deviation, and quantiles. Vertical red lines indicate the onset of fever, defined here as 1.5°C above the diurnal baseline sustained for 2h. These data are the features provided to the classification algorithm.

Figure 4

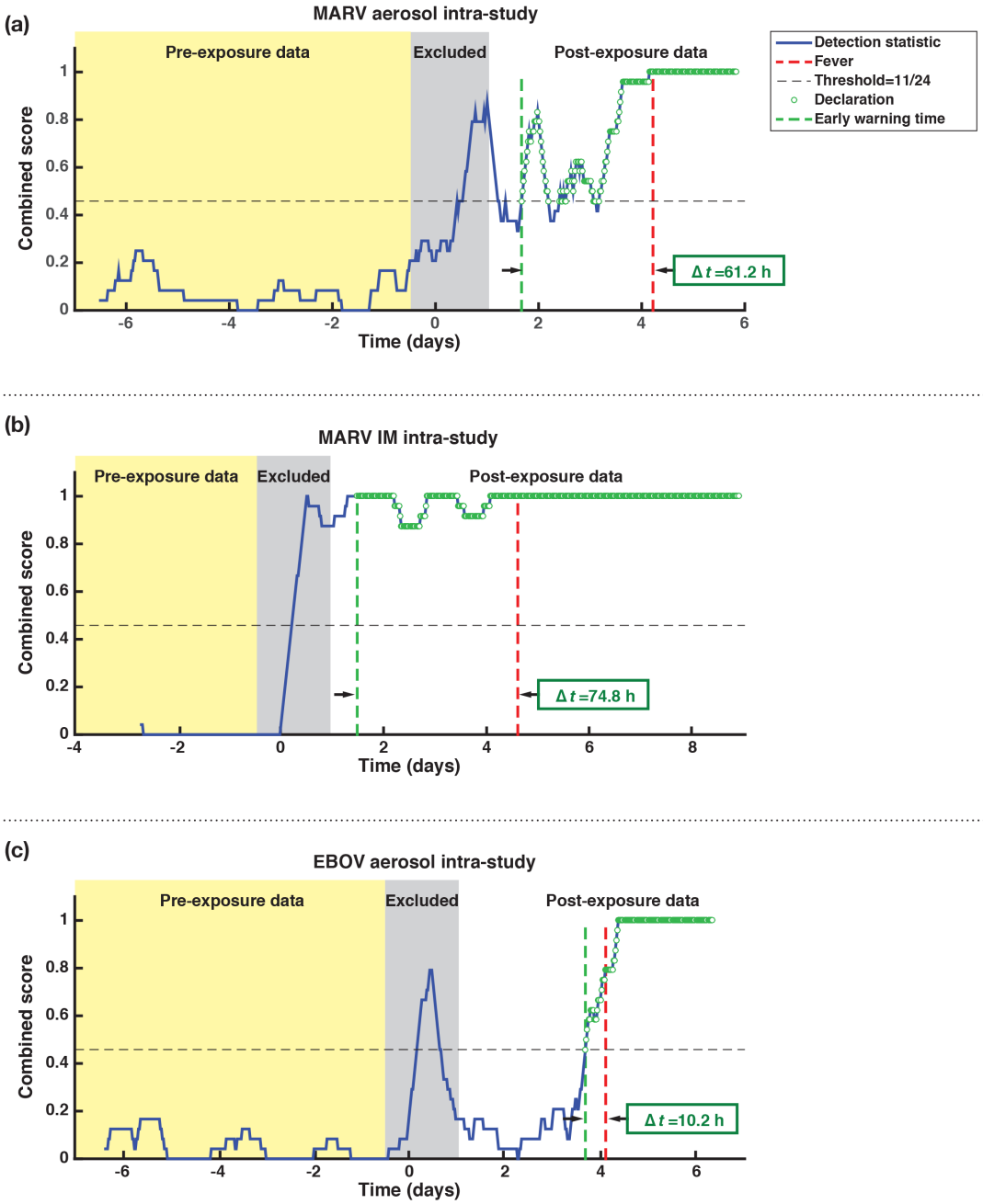

Figure 4: Representative single subject combined scores for the intra-study validations: (a) MARV aerosol, (b) MARV IM, and (c) EBOV aerosol. Scores are updated and plotted at 30 minute intervals and declarations (green triangles) are made when the score exceeds the m/n (11/24) threshold. Declarations before exposure (t=0) are false positives; scores after exposure below the dashed horizontal threshold line are false negatives. The time between the green and red vertical lines is the early warning time afforded by our algorithm. Note that data 12h before and 24h after exposure is disregarded due to animal anesthesia during exposure.

Figure 5

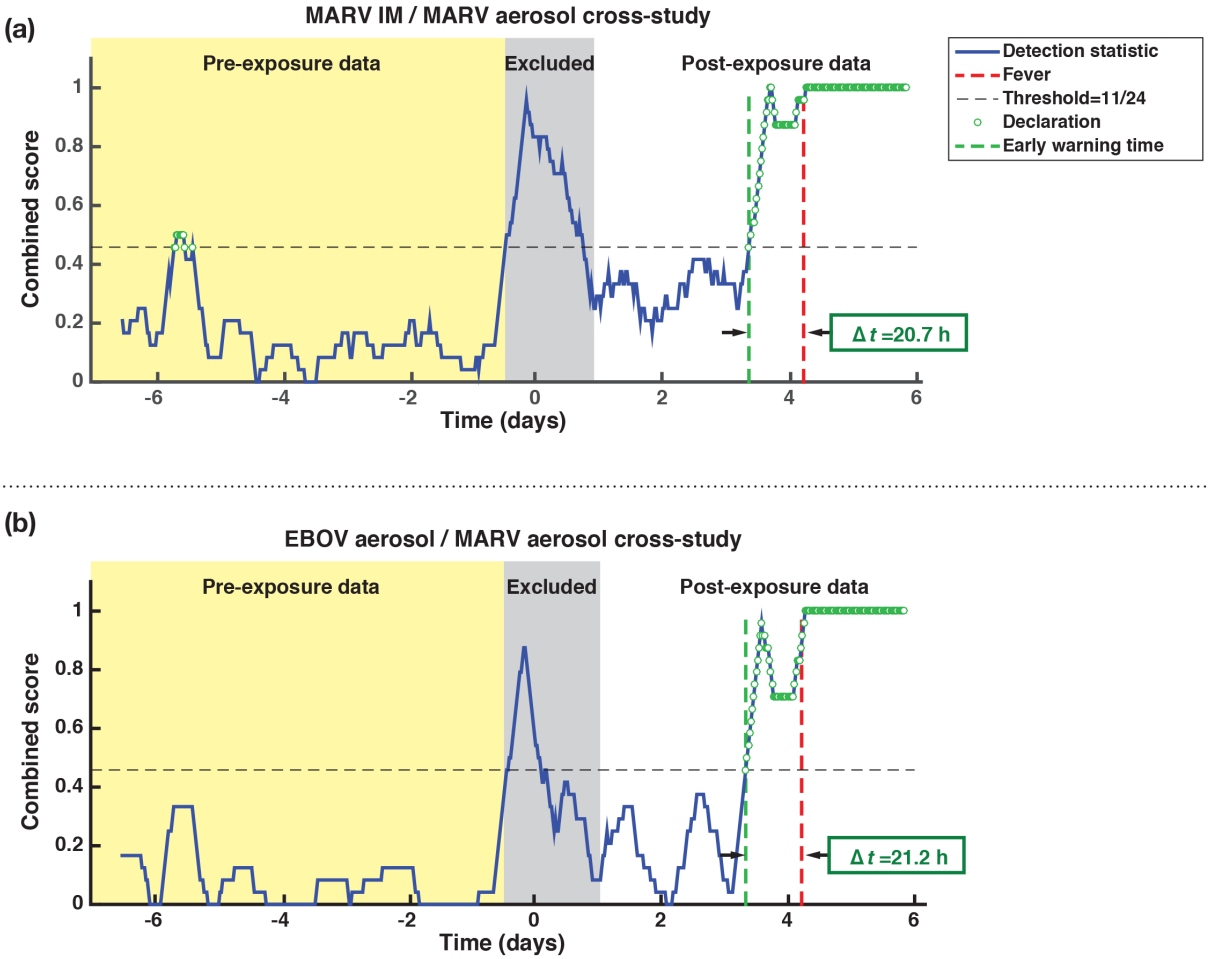

Figure 5: Examples of single subject algorithm outputs and declarations after false positive thresholding for two cross-study validations: (training set/testing set) (a) MARV IM/MARV aerosol, which use the same pathogen, and (b) EBOV aerosol / MARV aerosol, which holds the exposure route constant.

Figure 6

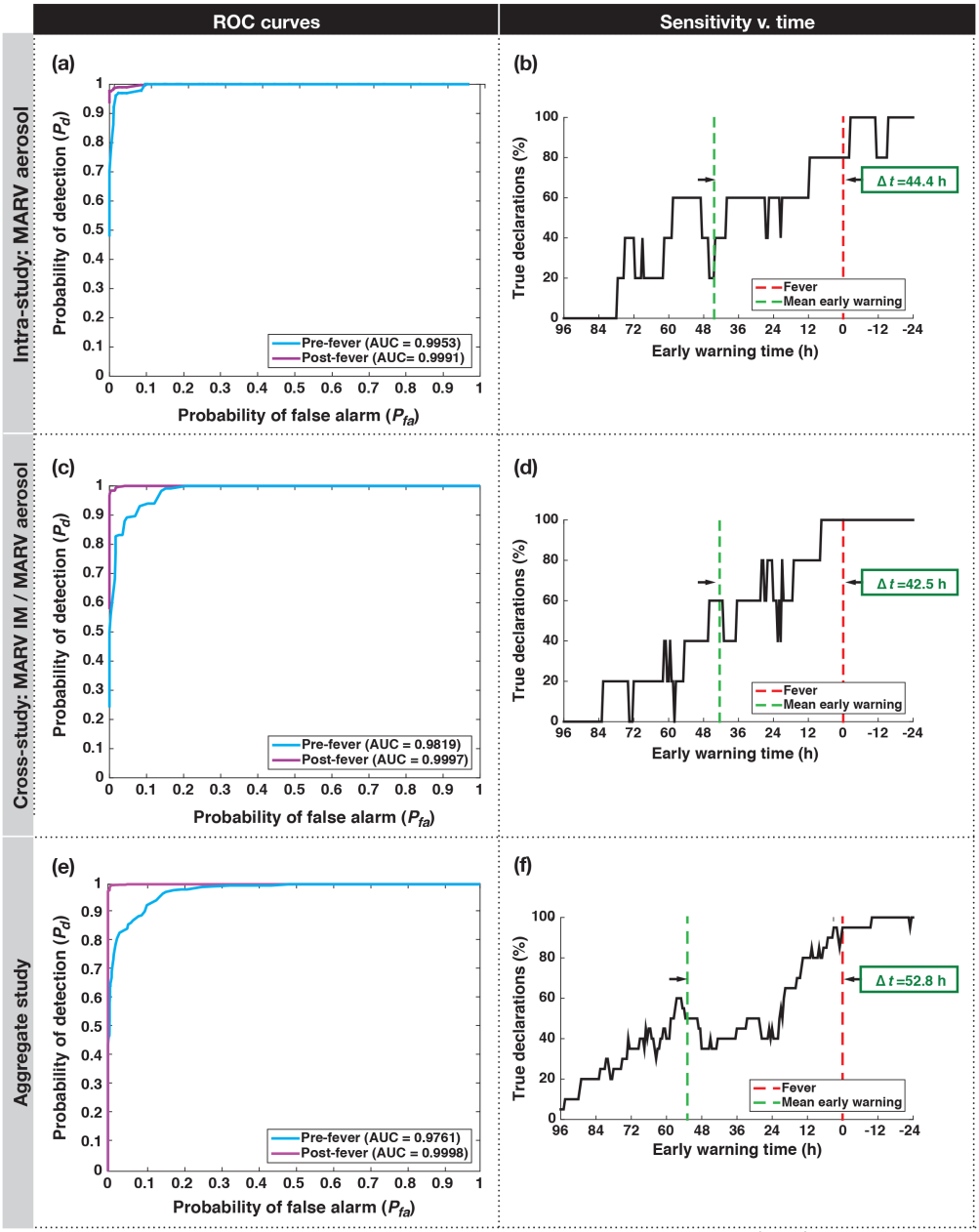

Figure 6: ROC curves and sensitivity vs. time plots for (a,b) MARV aero intra-study validation, (c,d) MARV IM/MARV aerosol cross-study, and (e,f) aggregated study validation tests. Nearly perfect algorithm performance is seen in the febrile prodrome, with only slightly lower performance during the incubation period. (b) , (d), and (f) show the percent of subjects correctly declared as "exposed" as a function of time before fever for the MARV aero intra-study (false detection rate P_{fa}<0.001), MARV IM/MARV aerosol cross-study (P_{fa}=0.04±0.01), and aggregated cross-study (P_{fa}<0.0005) validation tests, respectively. The green vertical lines represent the mean early warning time for the entire study.

Figure 7

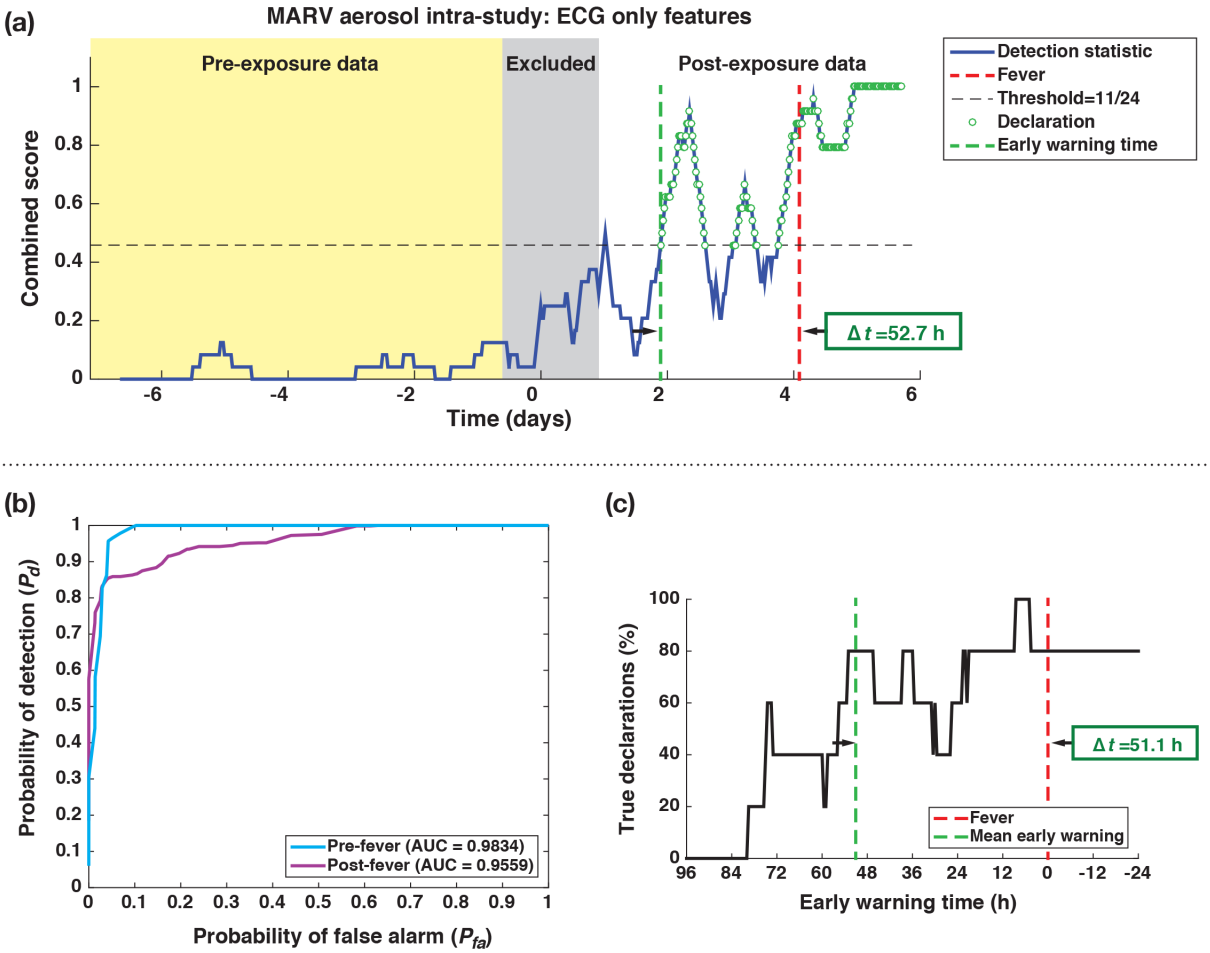

Figure 7: Using only ECG-derived features from the MARV aerosol study to train the model, one representative subject's (a) combined score output, (b) corresponding ROC curves for the entire study, and (c) percentage of correctly declared subjects versus early warning time. Since core body temperature is no longer available to the algorithm, model performance during the febrile prodrome is slightly worse than the pre-fever incubation period. Furthermore, while the overall AUC performance drops relative to the feature sets shown above, this limited set could be collected entirely using wearable monitoring devices.

Figure 8

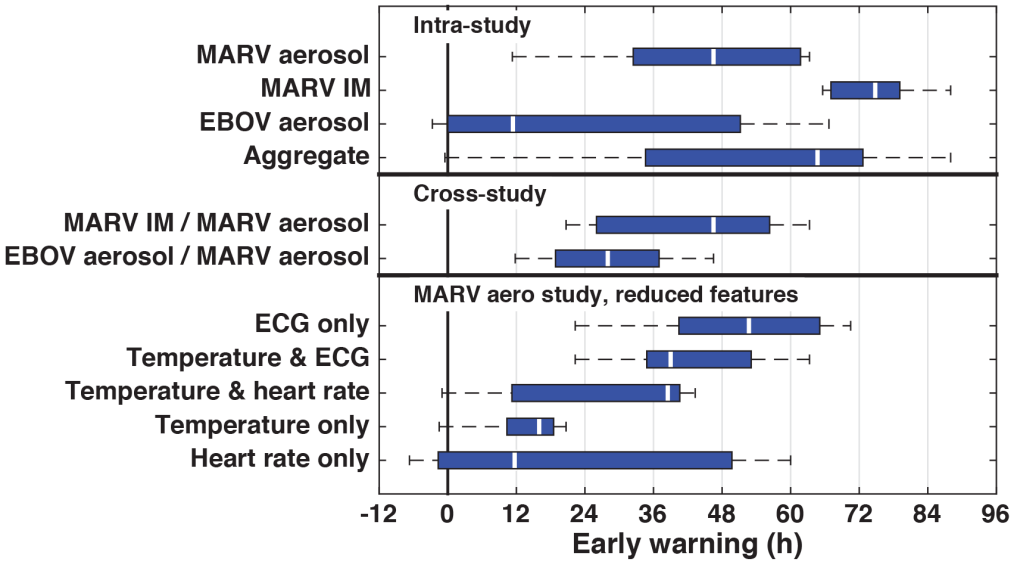

Figure 8: Box-and-whisker plot summarizing the early warning times from both intra- and cross-study validations. White vertical lines indicate the median value for each study, boxes show the first and third quartiles, and whiskers represent the largest and smallest value. Intra-study MARV tests using all available features give the largest early warning times, and the intra-study EBOV test showed the worst performance. Cross-study validations, including the aggregate study that considered all data over all studies, have very similar performance, suggesting algorithm robustness against virus strain and exposure routes. Reducing the feature set systematically degrades algorithm performance; the best performance is observed using all available ECG-derived features, and the worst performance when only heart rate is considered. This suggests that subtle electrophysiological features in the ECG signal (PR, QT, QRS intervals, etc.) are some of the most discriminating for our classification algorithm.